# Algoritmische risicotaxatie van recidive. Over de Oxford Risk of Recidivism tool (OXREC), ongelijke behandeling en discriminatie in strafzaken

# Algoritmische risicotaxatie van recidive

## Over de Oxford Risk of Recidivism tool (OXREC), ongelijke behandeling en discriminatie in strafzaken

Gijs van Dijck[1]

**Dit artikel verkent het risicotaxatie-instrument OXREC, dat door de reclassering wordt gebruikt en inmiddels ook de rechtszaal heeft bereikt. De conclusie is dat het wetenschappelijk onderzoek waarop het instrument is gebaseerd relevant is, maar de kwaliteit van het voorspellingsmodel vanuit praktisch oogpunt twijfelachtig. Met de introductie van OXREC lijken de Reclassering Nederland en ook rechters zich te laten leiden door voorspellingen die geregeld niet accuraat zijn, niet beter presteren dan reeds bestaande voorspellingsmodellen en waarbij ongelijke behandeling op basis van ras, sociale klasse of andere sociale ongelijkheden op de loer ligt.**

### 1. Inleiding

Risicotaxatie-instrumenten staan in de belangstelling van de media, vooral wanneer blijkt dat er mogelijk sprake is van etnisch profileren. Een recent voorbeeld is het *Systeem Risico Indicatie* (SyRI), dat de overheid gebruikte om fraude te bestrijden op onder meer het terrein van uitkeringen, toeslagen en belastingen. De rechter zette er een streep door: het was in strijd met het recht op privéleven zoals neergelegd in artikel 8 van het Europees Verdrag voor de Rechten voor de Mens (EVRM).[2] Een nog recenter voorbeeld betreft de Belastingdienst, die tussen 2012 en 2015 belastingplichtigen voor extra controle zouden hebben geselecteerd mede op basis van het hebben van een dubbele nationaliteit.[3] En er zijn mogelijk verschillende andere voorbeelden van etnisch profileren door de Nederlandse overheid.[4] Deze bijdrage bespreekt er een van, te weten risicotaxatie-instrumenten die de kans op recidive trachten te voorspellen.

Er is de nodige aandacht voor het gebruik van risicotaxatie-instrumenten in relatie tot recidive. Zo bestaat er een vracht aan onderzoek naar het voorspellen van recidive.[5] Deze studies richten zich doorgaans op een hoeveelheid factoren, zoals medische, strafrechtelijke en demografische gegevens, waarmee getracht wordt recidive van bepaalde misdrijven of delicten te voorspellen, in het bijzonder geweldsdelicten. Er is veel kritiek op dergelijke instrumenten: (1) ze meten niet precies wat ze beogen te meten, met veel foutmarges, (2) er is niet altijd overtuigend bewijs dat ze tot betere uitkomsten leiden dan wanneer mensen de risicotaxatie uitvoeren en (3) ze veroorzaken of vergroten ongelijke behandeling en discriminatie.[6] Deze problemen worden in een recente publicatie in dit tijdschrift op heldere wijze uitgelegd.[7]

Een bekend voorbeeld van een risicotaxatie-instrument is COMPAS. COMPAS, dat in verschillende staten in de VS wordt gebruikt, is omstreden om de hierboven genoemde redenen.[8] In de beroemde *Loomis/Wisconsin*-zaak werd het punt van nauwkeurigheid en ongelijke behandeling aan de orde gesteld, waar nog bij kwam dat het algoritme en de data waarop het algoritme getraind is niet of in beperkte mate inzichtelijk waren voor de verdachte (en de rechter).[9] De verdachte kreeg nul op het rekest en werd veroordeeld tot zes jaar gevangenisstraf, mede omdat het recidiverisico door COMPAS als hoog werd ingeschat.

In Nederland heeft een met COMPAS vergelijkbaar risicotaxatie-instrument, genaamd OXREC, recent haar intrede gedaan. Het instrument is gebaseerd op weten-

**De data waarop het algoritme getraind is was niet of in beperkte mate inzichtelijk voor de verdachte (en de rechter)**

schappelijk onderzoek onder een steekproef van veroordeelden. Aanvankelijk is het onderzoek uitgevoerd onder Zweedse veroordeelden,[10] hetgeen tevens heeft geleid tot een online tool waarmee het recidiverisico kan worden ingeschat.[11] Later is het instrument gevalideerd in de Nederlandse context aan de hand van data van het WODC, het Ministerie van Justitie en Veiligheid, het CBS en de RISc Database van de Nederlandse reclassering. Publicatie van laatstgenoemde studie vond plaats in 2019.[12] Ook op basis van dit onderzoek is een online tool gemaakt.[13]

Gelet op de ophef over risicotaxatie-instrumenten dook ik dieper in het instrument. Ik vroeg mij in dit verband af of het instrument wordt gebruikt, wat het meet en of er sprake is van etnisch profileren. Voor zover ik weet is er nog geen wetenschappelijke literatuur op dit punt. Ik doe hieronder verslag van mijn bevindingen.

## 2. Gebruik van OXREC

Een eerste vaststelling is dat OXREC sinds de publicatie in 2019 wordt gebruikt in de praktijk, met name door Reclassering Nederland (hierna: Reclassering). De Reclassering is een zelfstandige organisatie met als taak het toezien op de uitvoering van werkstraffen, het diagnosticeren van verdachten, het adviseren van rechters en officieren van justitie, en het toezichthouden op daders en verdachten. Het gebruik van OXREC gebeurt in het kader van de invulling van de tweede taak: diagnose en advies aan rechters en officieren van justitie. Afgaande op het onderzoek[14] en de online tool,[15] vult de reclasseringsambtenaar informatie in van de persoon voor wie het recidivegevaar wordt voorspeld. Het gaat om informatie zoals geslacht, leeftijd, detentieduur, relatiestatus (single/overig), opleidingsniveau, inkomen, alcoholgebruik, drugsgebruik en psychische aandoening.

OXREC is niet het enige instrument dat de Reclassering gebruikt om recidive te voorspellen. Een ander instrument betreft 'Recidive Inschattingsschalen' (RISc). Reeds in 2009 is onderzocht hoe RISc presteert ten aanzien van het voorspellen van recidive. Het onderzoek concludeerde dat RISc acceptabele voorspellingen geeft met betrekking tot algemene recidive en dat, hoewel het niet is bedoeld om specifiekere vormen van recidive te voorspellen, RISc ook acceptabele uitkomsten geeft voor de voorspelling van ernstiger vormen van recidive.[16] Het onderzoek richtte zich echter vooral op de vraag 'of er optimale grenswaarden bestaan op basis waarvan zoveel mogelijk daders goed kunnen worden geclassificeerd'; het was aan het beleid 'om te bepalen wat een acceptabel percentage vals positieven en vals negatieven is'.[17] Het is mij niet duidelijk waarom de vals positieven en vals negatieven niet zijn onderzocht. Was dat omdat dit niet in de opdracht stond en de onderzoekers er dus niet voor betaald kregen? Zo ja, waarom was het geen onderdeel van de opdracht? Zo nee, wat is dan de reden? Zoals ook

**Auteur**

**1.** Prof. mr. G. van Dijck is verbonden aan het Maastricht Law & Tech Lab, Universiteit Maastricht. Met dank aan Catalina Goanta, Matthias van der Haegen, Roland Moerland, Jerry Spanakis en Johan van Soest voor hun opmerkingen. Fouten en omissies komen, zoals gebruikelijk, voor rekening van de auteur.

**Noten**

**2.** Rb. Den Haag 5 februari 2020, ECLI:NL:RBDHA:2020:865 (*SyRi*).

**3.** 'Belastingdienst geeft toe: toch sprake van etnisch profileren' www.rtlnieuws.nl/nieuws/artikel/5117616/belastingdienst-toeslagen-profileren-nationaliteit; 'Belastingdienst erkent: toch sprake van etnisch profileren', www.trouw.nl/economie/belastingdienst-erkent-toch-sprake-van-etnisch-profileren~b91d1a45 (laatst geraadpleegd op 17 juni 2020).

**4.** 'Nationale Ombudsman ziet etnisch profileren in alle lagen van de overheid' www.nrc.nl/nieuws/2020/05/14/nationale-ombudsman-ziet-etnisch-profileren-in-alle-lagen-van-de-overheid-a3999711 (laatst geraadpleegd 17 op juni 2020); Ashley Terlouw, 'Gebruik van etniciteit in risicoprofielen Marechaussee is discriminatie', *NJB* 2020/764, afl. 12, p. 832-836.

**5.** Bijv. Bernard E. Harcourt, 'Risk as a Proxy for Race: The Dangers of Risk Assessment', *Federal Sentencing Reporter* 2015/4, p. 237-243 (overzicht van instrumenten); James McGuire, 'Minimising harm in violence risk assessment: Practical solutions to ethical problems?', *Health, Risk & Society* 2004/4, p. 327-345. Zie voorts onder meer de hieronder aangehaalde bronnen, alsook Eric Blaauw & Stefan Bogaerts & Marinus Spreen, 'Risicotaxatie in de Nederlandse rechtspraktijk: op naar een best practice', *Expertise en Recht* 2019/2, p. 71-77 voor verdere verwijzingen, ook naar studies over risicotaxaties in andere toepassingsgebieden dan alleen recidive.

**6.** Bijv. Sonja B. Starr, 'Evidence-Based Sentencing and the Scientific Rationalization of Discrimination', *Stanford Law Review* 2014, p. 803-872; Julia Dressel & Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 2018/1; Jennifer L. Skeem & Christopher Lowenkamp, 'Risk, Race, and Recidivism: Predictive Bias and Disparate Impact', *Criminology* 2016/4, p. 680-712; Danielle Kehl, Priscilla Guo & Samuel Kessler, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing', 2017; Danielle Kehl, Priscilla Guo & Samuel Kessler, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing', 2017; Aleš Završnik, 'Algorithmic justice: Algorithms and big data in criminal justice settings', *European Journal of Criminology* 2019, p. 1-20; Johannes Bijlsma, Floris Bex & Gerben Meynen, 'Artificiële intelligentie en risicotaxatie', *NJB* 2019/2778, afl. 44, p. 3313-3319; Joke Harte, 'Recidive inschatten met behulp van een empirisch model', *NJB* 2017/1799, afl. 33, p. 2386-2389. Niet alle studies wijzen op ongelijke behandeling, zie bijv. Jennifer L. Skeem & Christopher Lowenkamp, 'Risk, Race, and Recidivism: Predictive Bias and Disparate Impact', *Criminology* 2016/4, p. 680-712 (zwarte veroordeelden hebben een grotere kans op veroordeling, maar dit lijkt niet gerelateerd aan *bias*).

**7.** Johannes Bijlsma, Floris Bex & Gerben Meynen, 'Artificiële intelligentie en risicotaxatie', *NJB* 2019/2778, afl. 44, p. 3313-3319.

**8.** Bijv. Sandra G. Mayson, 'Bias in, Bias out', *Yale Law Journal* 2019/8, p. 2218-2301. Zie echter Cynthia Rudin, Caroline Wang & Beau Coke, 'The Age of Secrecy and Unfairness in Recidivism Prediction', *Harvard Data Science Review* 2020/1 (COMPAS-voorspellingen niet noodzakelijk gebaseerd op ras); Tim Brennan, William Dieterich & Beate Ehret, 'Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System', *Criminal Justice and Behavior* 2008/1, p. 21-40 (COMPAS presteert beter dan andere modellen). Zie overigens Elaine Angelino e.a., 'Learning Certifiably Optimal Rule Lists for Categorical Data', *Journal of Machine Learning Research* 2018, p. 1-78, waarin een voorspellingsmodel is ontwikkeld dat volledig interpreteerbaar is.

**9.** *Loomis/Wisconsin*, 881 N.W.2d 749 (Wis. 2016).

**10.** Seena Fazel e.a., 'Identification of low risk of violent crime in severe mental illness with a clinical prediction tool (Oxford Mental Illness and Violence tool [OxMIV]): a derivation and validation study', *The Lancet Psychiatry* 2017/6, p. 461-468.

**11.** https://oxrisk.com/oxrec/> (laatst geraadpleegd op 17 juni 2020).

**12.** Seena Fazel e.a., 'Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec)', *Scientific Reports* 2019.

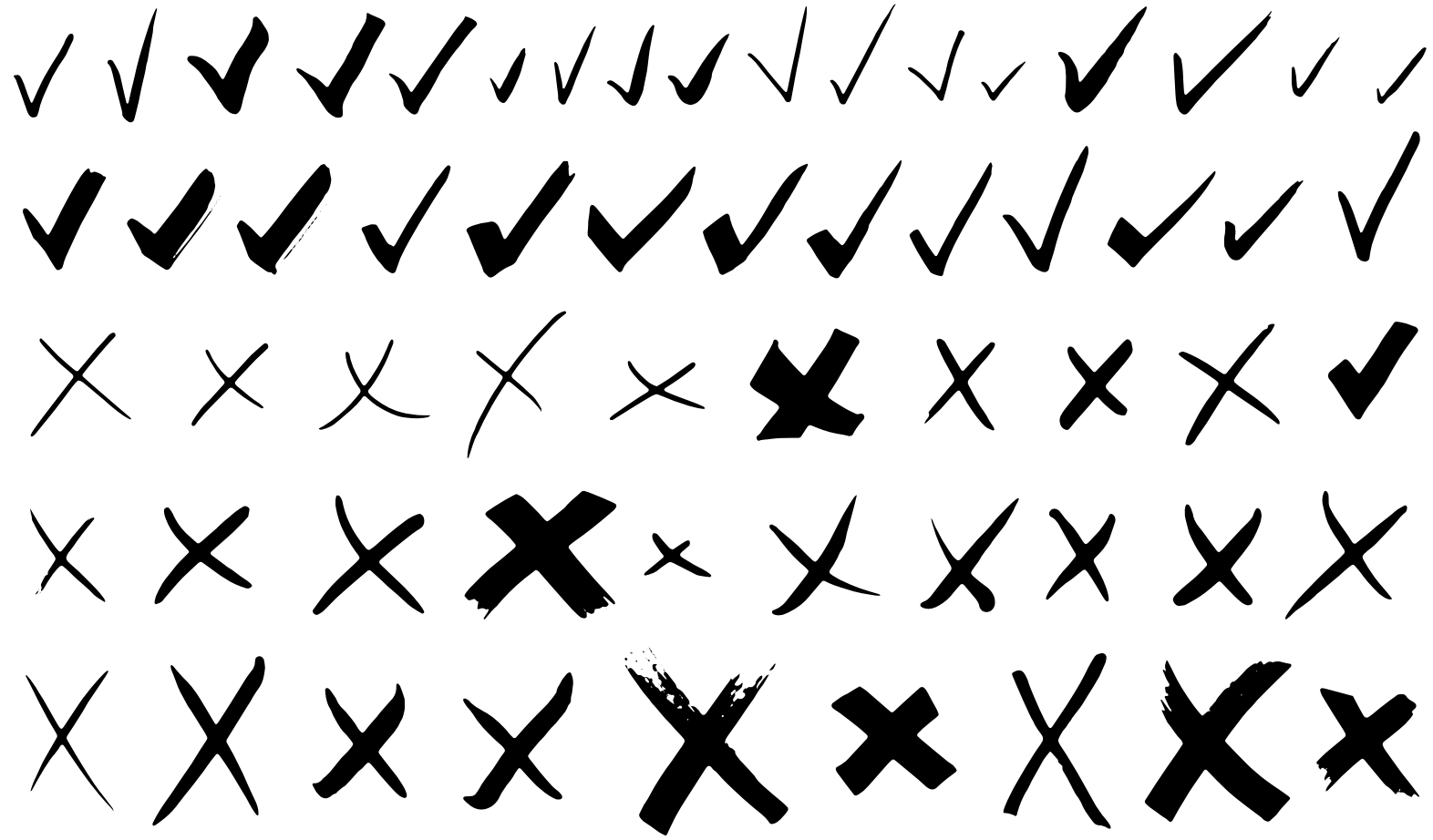
**13.** https://oxrisk.com/oxrec-nl-2-backup/ (laatst geraadpleegd op 17 juni 2020).

**14.** Seena Fazel e.a., 'Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec)', *Scientific Reports* 2019.

**15.** https://oxrisk.com/oxrec-nl-2-backup/ (laatst geraadpleegd op 17 juni 2020).

**16.** L.M. van der Knaap & L.D. Alberda, *De predictieve validiteit van de Recidive Inschattingsschalen (RISc)*, WODC 2009, p. 39.

**17.** L.M. van der Knaap & L.D. Alberda, *De predictieve validiteit van de Recidive Inschattingsschalen (RISc)*, WODC 2009, p. 30.



© Shutterstock

hieronder zal blijken, is bedoelde informatie essentieel om de praktische toepasbaarheid te beoordelen.

Terug naar OXREC. Daarover valt op de website van de Reclassering te lezen:

> **'Inschatting recidiverisico met OXREC**
> Onderdeel van de RISC is de OXREC. De OXREC ondersteunt de reclasseringswerker bij het maken van een gestructureerd professioneel oordeel; het reclasseringsoordeel. De OXREC is een actuarieel risicotaxatie-instrument van de University of Oxford waarmee op basis van statische en dynamische factoren een statistische berekening wordt gemaakt van het risico op zowel algemene en geweldsrecidive. In 2017 is de OXREC door de University of Oxford gevalideerd voor Nederland op basis van data van het CBS, het WODC en de 3RO.
> Actuariële risicotaxatie-instrumenten, en dus ook de OXREC, zijn gebaseerd op de samenhang tussen kenmerken van personen en recidivegegevens. Het geeft een indicatie van het recidiverisico van groepen personen met deze kenmerken. Naast de OXREC maakt de reclassering een inschatting van het recidiverisico in een specifieke casus, een inschatting over het individu; het reclasseringsoordeel. Dit reclasseringsoordeel- het gestructureerd professioneel oordeel- kan dus afwijken van de uitkomst van de OXREC.'[18]

Vermoedelijk vult de Reclassering informatie in op een vergelijkbare wijze als in de tool die online beschikbaar is gemaakt,[19] met als resultaat een risicotaxatie in drie categorieën ('laag', 'gemiddeld', 'hoog') met ook de onderliggende percentages (bijvoorbeeld 15% kans op recidive na één jaar, 22% kans na twee jaar). De Reclassering weegt naar eigen zeggen vervolgens de resultaten die OXREC produceert. Ik kon niet achterhalen hoe dat gebeurt, en welk gewicht aan OXREC wordt gegeven in relatie tot de andere bronnen van informatie (RISc, inschatting door reclassering in specifieke casus en van het individu). Hoe dan ook, afgaand op de beschikbare informatie is het uiteindelijk de Reclassering die een inschatting maakt van het risico. Daarmee is er een 'human in the loop': men gaat, in ieder geval in theorie, maar ook in de praktijk (zie hieronder), niet blind af op de door het door de tool geproduceerde uitkomst.

Sinds de verschijning van het onderzoek in 2019 zijn er, op het moment van schrijven van dit artikel, negen gerechtelijke uitspraken gepubliceerd waarin OXREC expliciet wordt genoemd. In deze uitspraken wordt ingegaan op de uitkomst van de OXREC-risicotaxatie,[20] soms aangevuld met de waarde waarop de score is gebaseerd,[21] met de inschatting door de Reclassering dan wel vermelding of de Reclassering het eens is met de taxatie.[22] Gelet op de gevonden rechtspraak lijkt het erop dat de rechter alleen de risicotaxatie krijgt in termen van 'laag', 'gemiddeld' en 'hoog', maar niet de onderliggende percentages. Dit is begrijpelijk, omdat de percentages onterecht de indruk kunnen wekken dat de taxaties precies zijn.

Slechts in twee van bovenstaande gevallen maakt de rechtbank een opmerking over het gebruik OXREC. In een

**Gelet op de gevonden rechtspraak lijkt het erop dat de rechter alleen de risicotaxatie krijgt in termen van 'laag', 'gemiddeld' en 'hoog'**

van de twee zaken gaat het om een herhaling van de risicotaxatie door de Reclassering.[23] De andere casus betreft een situatie waarin de Reclassering, vanwege het COVID-19-virus, de inhoud van het reclasseringsadvies niet met de verdachte heeft kunnen bespreken.[24] Verder stelt de rechtbank in twee gevallen dat het advies niet wordt gevolgd, maar dit betreft het advies van de Reclassering als geheel en niet alleen het OXREC-rapport.[25]

## 3. Hoe goed voorspelt OXREC?

### *3.1 Welke data gaan er in het voorspellingsmodel?*

OXREC wordt dus gebruikt. Maar hoe goed voorspelt het instrument? Het antwoord op deze vraag is in belangrijke mate afhankelijk van de data en de kwaliteit ervan. Het onderzoek waarop de tool is gebaseerd bestond uit een analyse van 9072 vrijgelaten veroordeelden (6% vrouw) en 6329 personen met een proeftijd (12% vrouw) in de periode 2011-2012. De personen hadden mediane leeftijden van respectievelijk 30 en 34. Van de veroordeelden ging 16% binnen twee jaar opnieuw de fout in (althans, voor zover geregistreerd) op een gewelddadige wijze, en 44% op gewelddadige of niet-gewelddadige wijze.

Het in Nederland uitgevoerde onderzoek volgde zo veel mogelijk de Zweedse studie waar het gaat om de verzamelde variabelen en gehanteerde definities.[26] Zoals hierboven aangegeven, is het onderzoek gebaseerd op een reeks variabelen, waaronder geslacht, leeftijd, deten-

**18.** www.reclassering.nl/over-de-reclassering/wat-wij-doen/risc (laatst geraadpleegd op 17 juni 2020).

**19.** https://oxrisk.com/oxrec-nl-2-backup/ (Nederland), https://oxrisk.com/oxrec/ (Zweden) (laatst geraadpleegd op 17 juni 2020).

**20.** Rb. Den Haag 23 juli 2019, ECLI:NL:RBDHA:2019:7431 ('Het risico op geweldsrecidive wordt door OXREC en de reclassering als gemiddeld ingeschat'); Rb. Den Haag 23 juli 2019, ECLI:NL:RBDHA:2019:7430 ('Het risico op geweldsrecidive wordt door OXREC en de reclassering als hoog ingeschat'; Engelstalige versie: ECLI:NL:RBDHA:2019:10647).

**21.** Rb. Gelderland 20 maart 2020, ECLI:NL:RBGEL:2020:1958 ('OXREC geeft aan dat de risico's binnen 2 jaar gemiddeld is. De gemiddelde score is gebaseerd op zijn delict geschiedenis, financiën en intieme relaties. Wanneer het gaat om de inschatting van het indexdelict worden de risico's laag-gemiddeld geschat op langere termijn').

**22.** Rb. Overijssel 30 april 2020, ECLI:NL:RBOVE:2020:1636 ('Uit een risicotaxatie met behulp van het instrument OXREC volgt dat de kans op recidive als hoog moet worden ingeschat. Het risico op geweldsdelicten is laag. De rapporteur is het eens met de uitkomsten van deze risicotaxatie'); Rb. Noord-Holland 24 mei 2019, ECLI:NL:RBNHO:2019:4795 (rechtbank citeert het reclasseringsrapport: 'Kijkend naar de score op de OXREC in combinatie met het delictverleden en de geconstateerde problematiek wordt door rapporteur de kans op recidive met letselschade ingeschat als hoog. De [verdachte] is bekend met herhaaldelijk grensoverschrijdend (delict) gedrag (agressie), voortkomend uit forse psychische problematiek, waarbij er nauwelijks sprake lijkt van inzicht in zijn eigen handelen'); Rb. Limburg 6 februari 2019, ECLI:NL:RBLIM:2019:3082 (rechtbank citeert het reclasseringsrapport: 'Met behulp van de RISC en OXREC (risicotaxatie die gehanteerd wordt door de reclassering) schatten wij het recidiverisico op dit moment in op gemiddeld. Bij betrokkene is sprake van actuele middelenproblematiek, thans deels in remissie door een ontwenningsmiddel (Refusal). Daarnaast is sprake van persoonlijkheidsproblematiek en een chronische post-stressstoornis. Beschermend is de huidige schematherapie, het medicatievoorschrift en de onderbewindstelling. Met name is het van belang dat betrokkene medicatietrouw is en zich onthoudt van middelengebruik. Mocht het huidige kader thans wegvallen dan schatten we de kans op algemeen risico in op hoog, omdat betrokkene zich niet staande zal kunnen houden zonder behandeling en begeleiding. De kans op gewelddadig gedrag schatten we dan in op gemiddeld'); Rb. Noord-Holland 6 juni 2019, ECLI:NL:RBNHO:2019:4944 ('De reclassering heeft een recidivekans berekend op basis van de Oxrec, de Static-R-99 en de Stable 2007 waarbij een lage kans op recidive naar voren komt. Vanwege de ernst, de omvang en de duur van het misbruik schat de reclassering de kans op recidive toch iets hoger in'); Rb. Midden-Nederland 18 maart 2019, ECLI:NL:RBMNE:2019:1108 ('Het risico op recidive wordt op basis van de OXREC geschat als hoog. Dit komt mede door de financiële situatie van het gezin, waarbij sprake is van een gezamenlijke uitkering en een forse schuldenlast. Er zijn aanwijzingen voor een deviant netwerk, beperkte copingvaardigheden voor de verschillende problemen en een beperkte intelligentie'); Rb. Amsterdam 19 februari 2019, ECLI:NL:RBAMS:2019:1076 ('Het recidiverisico is ingeschat op basis van onder andere de OXREC (Oxford Risk of Recidivism tool), een actuarieel instrument dat het risico op algemene- en geweldsrecidive meet. Op basis van alle risicofactoren, beschermende factoren en de gebruikte wetenschappelijke verdiepende instrumenten komt de reclasseringswerker tot een inschatting van het risico op recidive dat wordt ingeschat als hoog').

**23.** Rb. Limburg 6 februari 2019, ECLI:NL:RBLIM:2019:3082 ('De risico inschatting is met professionele ondersteuning matig. Zonder ondersteuning in de maatschappij is de risico inschatting matig/hoog').

**24.** Rb. Overijssel 30 april 2020, ECLI:NL:RBOVE:2020:1636 ('De rechtbank is zich terdege bewust van de beperkingen, verband houdende met maatregelen wegens de uitbraak van het Covid-19 virus, waardoor de rapporteur van GGZ IrisZorg verdachte voorafgaand aan het opmaken van de rapportage niet zelf heeft kunnen spreken. Evenmin heeft de rapporteur het reclasseringsadvies vóór de zitting op verantwoorde wijze met verdachte kunnen bespreken. De rechtbank ziet gezien de inhoud van de rapportage echter geen redenen om te oordelen dat het advies onzorgvuldig tot stand is gekomen of om aan de inhoud ervan te twijfelen. Gelet op de reeds aanwezige informatie heeft rapporteur zijn advies op verantwoorde wijze daarop kunnen baseren. De rechtbank neemt hierbij in aanmerking dat het advies om, in geval van een veroordeling, een onvoorwaardelijke ISD-maatregel op te leggen degelijk is onderbouwd en goed is gemotiveerd, onder meer door middel van een overzichtelijke weergave met analyse van alle risico- en delict gerelateerde factoren en een weergave van alle reclasserings- en hulpverleningscontacten in het (ook recente) verleden. Het reclasseringsadvies kan dan ook worden aangemerkt als een in artikel 38m Sr bedoeld advies').

**25.** Rb. Den Haag 23 juli 2019, ECLI:NL:RBDHA:2019:7431 ('De rechtbank volgt dit advies [van de reclassering] niet'); Rb. Den Haag 23 juli 2019, ECLI:NL:RBDHA:2019:7430 ('De rechtbank volgt dit advies [van de reclassering] niet'; Engelstalige versie: ECLI:NL:RBDHA:2019:10647).

**26.** Zie https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-37539-x/MediaObjects/41598_2018_37539_MOESM1_ESM.pdf voor een overzicht en definities van de variabelen die zijn gebruikt en in hoeverre zij overeenkomen/afwijken van de Zweedse studie (laatst geraadpleegd op 17 juni 2020).

**Verschillende van deze variabelen en indicatoren zouden als mogelijk discriminerend kunnen worden aangemerkt**

tieduur, relatiestatus (single/overig), opleidingsniveau, inkomen, alcoholgebruik, drugsgebruik en psychische aandoening.[27] Dit is tevens de input waarom wordt gevraagd in de online tool.[28] De variabele *Deprivation* is interessant vanuit het perspectief van etnisch profileren. Deze variabele is samengesteld op basis van een vijftal indicatoren, te weten de postcode, ontvangst van bijstandsuitkering (*welfare recipiency*), werkloosheid, laag opleidingsniveau, criminaliteitscijfers en mediaan inkomen.

Verschillende van deze variabelen en indicatoren zouden als mogelijk discriminerend kunnen worden aangemerkt. Postcode, geslacht, leeftijd, opleidingsniveau en inkomen kunnen werkelijk goede voorspellers zijn, maar zij kunnen ook (deels) direct of indirect indicatoren zijn voor etnische profilering. Indien personen in bepaalde bevolkingsgroepen, met een bepaalde etniciteit, in een bepaalde sociale klasse etc. een grotere kans hebben om te worden opgepakt en te worden veroordeeld dan anderen, zullen genoemde variabelen deze informatie reflecteren.[29]

### *3.2 Het voorspellingsmodel*

Het voorspellingsmodel zal mogelijk discriminerende patronen herkennen, waardoor bepaalde groepen (verder) zullen worden benadeeld. Dit blijkt bijvoorbeeld wanneer ik waardes ging invullen in de online tool.[30] De mogelijke ongelijkheid is overigens niet zozeer een probleem veroorzaakt door het algoritme, maar is het gevolg van de data waarmee het algoritme wordt gevoed.

Voorts is er de vraag waar een op een algoritme gebaseerd voorspellingsmodel mee wordt vergeleken. Doorgaans vormen menselijke beslissers de referentiegroep, maar ook mensen maken fouten en zijn bewust of onbewust bevooroordeeld.[31] Bovendien is het empirisch bewijs niet eenduidig. Aan de hand van wetenschappelijk bewijs kan worden beweerd dat statistische modellen beter voorspellen dan mensen, maar er is ook bewijs waaruit het tegenovergestelde volgt.[32] Een en ander dient te worden onderzocht in de specifieke context waarin OXREC wordt toegepast. Daarnaast kan het niet-meenemen van discutabele variabelen resulteren in voorspellingsmodellen die *meer* discrimineren dan een model *met* de variabelen (bijvoorbeeld hogere risicotaxatie voor vrouwen bij weglaten geslacht-variabele).[33] Bovendien kunnen vele historische variabelen potentieel discriminatoir zijn.[34]

Verder zijn er aanwijzingen dat voorspellingsmodellen met een breed palet aan variabelen beter presteren dan die met een beperkt aantal variabelen,[35] maar er is ook bewijs dat voorspellingsmodellen met slechts enkele variabelen en met traditionele analyses (bijvoorbeeld logistische regressieanalyse) net zo goed of zelfs beter presteren dan modellen ontwikkeld met geavanceerde *deep learning* of *neural network*-technieken.[36] Er is dan ook nog veel onduidelijk over de kwaliteit van dergelijke voorspellingsmodellen. Een en ander zal doorgaans in een specifieke context, bijvoorbeeld voor recidive in Nederland, of zelfs voor een specifiek delict of specifieke doelgroep, moeten worden onderzocht.

### *3.3 Kwaliteit van uitkomsten: statistisch perspectief*

De kwaliteit van de uitkomsten van het voorspellingsmodel kan vanuit statistisch perspectief op verschillende manieren worden beoordeeld.[37] Zogenaamde *Area Under Curve (AUC) percentages* kwantificeren de geschiktheid van het instrument ten aanzien van het onderscheiden van personen in een bepaalde risicogroep en personen die niet tot die risicogroep behoren. Een AUC van 1.0 doet dit perfect (geen vals alarm, geen vals negatieven), terwijl een score van 0.5 aangeeft dat het niet beter presteert dan het opgooien van een muntje. Daarbij is het tevens van belang om te weten wat belangrijk is: precies aangeven of iemand met een bepaald risico (bv. Recidiverisico = 'hoog') als zodanig door het instrument wordt herkend, of voorkomen dat personen onterecht in een bepaalde categorie (bijvoorbeeld Recidiverisico = 'hoog') worden geplaatst. Statistische evaluatiecriteria die hierbij kunnen helpen en die ook in de hier besproken studie zijn gebruikt, zijn:

- *Sensitiviteit*: de verhouding tussen het aantal personen dat positief scoort voor een bepaalde risicogroep en het totaal van de onderzochte personen die in die risicogroep vallen.
- *Specificiteit*: het percentage dat *niet* tot een bepaalde groep behoort en (terecht) als zodanig door het instrument wordt aangemerkt.
- *Positive predicted value* (PPV): de kans dat het instrument een vals alarm geeft.
- *Negative predicted value* (NPV): de kans dat het instrument 'stil' blijft waar het alarm had moeten afgaan.

Aan de hand van bepaalde drempels ('risk thresholds') kan een balans worden gevonden tussen bovenstaande evaluatiecriteria. Een lage drempel zal leiden tot het correct aanwijzen van degenen die in een bepaalde risicogroep thuishoren, maar ook tot het aanwijzen van veel personen die er niet in thuishoren (vals alarm). Omgekeerd geldt bij een hoge drempel dat de kans klein is dat personen in een risicogroep worden geplaatst waar ze eigenlijk niet thuishoren, maar tevens dat er meer personen die in een bepaalde risicogroep thuishoren niet als zodanig worden opgemerkt. Bij het voorspellen van recidive ligt het voor de hand in ieder geval veel gewicht toe te kennen aan het voorkomen van vals alarm in de risicocategorie 'hoog', zodat personen niet onterecht in de hoogste risicocategorie worden geplaatst.

De resultaten van de studie (zie Tabel 1)[38] laten AUC-scores zien van tussen de 0.67 en 0.69, met betrouwbaarheidsintervallen[39] tussen 0.65 en 0.70. Dit suggereert dat de kans op een juiste classificatie beter is dan het opgooien van een muntje (AUC=0.50), maar dat er nog altijd een 30%-35% kans is op vals alarm en/of stilte waar het alarm had moeten afgaan.

**Tabel 1: Samenvatting gekalibreerd model OXREC-studie met Nederlandse data**

From: Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec)

| | Risk threshold | Prevalence of reoffending | Sensitivity | Specificity | PPV | NPV | c-index (95% CIS) |
|---|---|---|---|---|---|---|---|
| Violent reoffending, 2 yr, prisoners | 10% | 16% | 91% (89-92) | 32% (31-34) | 20% (19-22) | 95% (94-96) | 0.68 (0.66-0.70) |
| | 30% | | 12% (10-14) | 94% (93-95) | 27% (23-32) | 85% (84-86) | |
| Any reoffending, 2 yr, prisoners | 30% | 44% | 90% (89-91) | 31% (30-33) | 51% (50-52) | 81% (78-82) | 0.69 (0.68-0.70) |
| | 50% | | 50% (48-52) | 74% (73-76) | 60% (58-62) | 65% (64-67) | |
| Violent reoffending, 2 yr, non-prisoners | 10% | 11% | 71% (66-74) | 59% (58-60) | 17% (15-18) | 95% (94-95) | 0.68 (0.65-0.70) |
| Any reoffending, 2 yr, non-prisoners | 30% | 28% | 54% (51-56) | 69% (68-71) | 40% (38-43) | 79% (78-81) | 0.67 (0.65-0.68) |

PPV = Positive predictive value; NPV = Negative predictive value. Note: the 30% (and 50%) threshold was not useful for non-prisoners, as very few had a predicted risk that exceeded this.

Gezien het belang van het voorkomen van vals alarm, zeker in gevallen waar de consequenties het grootst zijn, zoals bij een risico-inschatting van gewelddadige recidive, is het belangrijk om tevens te kijken naar de PPV's. Een inspectie hiervan levert op dat geen van de PPV's uitkomt boven de 60%, met soms percentages van 27% of zelfs 17%. OXREC geeft dus vaak vals alarm, soms zelfs in 70%-80% van de gevallen, wat betekent dat veroordeelden relatief vaak onterecht in een bepaalde risicogroep worden geplaatst. In het meest gunstige geval is er sprake van vals alarm in 40% van de gevallen. Dat is zorgwekkend en geen goed nieuws voor de (vele) veroordeelden met een onterecht 'hoog' voorspeld recidivegevaar. Wel moet worden opgemerkt dat het onduidelijk is of het voorspellingsmodel beter dan wel slechter presteert dan wanneer experts (mensen) voorspellen. Het zou kunnen dat zij (nog) slechter voorspellen dan het voorspellingsmodel.

## 4. Conclusie

De conclusie is drieledig:

1) Het wetenschappelijk onderzoek waarop OXREC is gebaseerd is relevant. Meer onderzoek is gewenst en nodig om te bezien of dan wel in hoeverre het mogelijk is om risicotaxatie te verbeteren.
2) De kwaliteit van het voorspellingsmodel is vanuit praktisch oogpunt twijfelachtig. Toepassing is alleen mogelijk indien men grote foutmarges accepteert. Het instrument geeft vaak vals alarm, maar het is tevens onduidelijk of menselijke voorspellers het beter doen. Om zicht te krijgen op het laatste, zou een prospectieve studie moeten worden uitgevoerd die de uitkomsten van het voorspellingsmodel voor een willekeurig gekozen groep veroordeelden vergelijkt met de voorspellingen van deskundigen die geen gebruikmaken van het model (en mogelijk met een derde groep: deskundigen

**27.** In het Zweedse onderzoek werd immigrantenstatus nog meegenomen als variabele, maar deze informatie bleek niet beschikbaar in de datasets gebruikt voor de Nederlandse studie.

**28.** https://oxrisk.com/oxrec-nl-2-backup/ (laatst geraadpleegd op 17 juni 2020).

**29.** Richard S. Frase, 'What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations?', *Crime and Justice: A Review of Research* 2009, p. 201-280.

**30.** https://oxrisk.com/oxrec/ (Zweden), https://oxrisk.com/oxrec-nl-2-backup/ (Nederland) (laatst geraadpleegd op 17 juni 2020). Zie voorts Derek W. Braverman e.a., 'OxRec model for assessing risk of recidivism: ethics', *The Lancet Psychiatry* 2016/9, p. 808-809.

**31.** In die zin zal ieder voorspellingsmodel doorgaans gedoemd zijn te falen, aangezien zij op het verleden zijn gebaseerd waar *bias* een gegeven is. Zie Sandra G. Mayson, 'Bias in, Bias out', *Yale Law Journal* 2019/8, p. 2218-2301.

**32.** Zie bijv. Sonja B. Starr, 'Evidence-Based Sentencing and the Scientific Rationalization of Discrimination', *Stanford Law Review* 2014, p. 803-872 en de daarin opgenomen verwijzingen (bewijs dat statistische modellen beter presteren dan experts is twijfelachtig) versus Stefanía Ægisdóttir e.a., 'The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Researchon Clinical Versus Statistical Prediction', *The Counseling Psychologist* 2006/3. Zie voorts Julia Dressel & Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 2018/1 (COMPAS presteert niet beter dan personen met nauwelijks tot geen strafrechtelijke expertise); Joke Harte, 'Recidive inschatten met behulp van een empirisch model', *NJB* 2017/1799, afl. 33, p. 386-388.

**33.** John Monahan & Jennifer L. Skeem, 'Risk Assessment in Criminal Sentencing', *Annual Review Clinical Psychology* 2016, p. 489-513.

**34.** Seena Fazel e.a., 'OxRec model for assessing risk of recidivism: ethics - Authors' reply', *The Lancet Psychiatry* 2016/9, p. 809-810.

**35.** Jennifer L. Skeem & Christopher Lowenkamp, 'Risk, Race, and Recidivism: Predictive Bias and Disparate Impact', *Criminology* 2016/4, p. 680-712.

**36.** Julia Dressel & Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 2018/1; James E. Johndrow & Kristian Lum, 'An algorithm for removing sensitive information: application-to race-independent recidivism prediction', *arXiv:1703.04957v1 [stat.AP]* 2017; Chris Baird e.a., 'A Comparison of Risk Assessment Instruments in Juvenile Justice', 2013, p. 134.

**37.** Voor een introductie, zie bijv. Johannes Bijlsma, Floris Bex & Gerben Meynen, 'Artificiële intelligentie en risicotaxatie', *NJB* 2019/2778, afl. 44, p. 3313-3319; Rajul Parikh e.a., 'Understanding and using sensitivity, specificity and predictive values', *Indian Journal Ophthalmol* 2008/1, p. 45-50.

**38.** Seena Fazel e.a., 'Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec)', *Scientific Reports* 2019.

**39.** Er kan met 95% zekerheid worden gezegd dat het werkelijke percentage binnen het interval valt.

## Het gebruik van OXREC creëert dan wel versterkt de *mogelijkheid* van ongelijke behandeling op basis van ras, sociale klasse of andere sociale ongelijkheden

die gebruik maken van zowel hun eigen expertise als-ook het voorspellingsmodel).

3) Het gebruik van OXREC creëert dan wel versterkt de *mogelijkheid* van ongelijke behandeling op basis van ras, sociale klasse of andere sociale ongelijkheden.[40] Indien de Reclassering en, in navolging hiervan, de rechter hier geen acht op slaat, is er een reële kans dat uitspraken worden gewezen die mede worden beïnvloed door een instrument dat personen van een bepaalde afkomst of met een bepaald etnisch profiel (nog verder) op achterstand zetten in het strafproces. Oplossingen zijn om de tool niet te gebruiken, of om variabelen gerelateerd aan een mogelijke ongelijke behandeling niet mee te nemen in de tool. Bij laatstgenoemde oplossing rijst wel de vraag wat dit doet met de kwaliteit van de uitkomsten van de tool. Onderzoek daarover is ambigu.

Het is twijfelachtig of het onderzoek al zodanig gevorderd is dat het in de praktijk kan worden gebruikt. Dit betekent overigens niet dat het gebruik leidt tot ongewenste uitkomsten. Dat is niet te bepalen. Zo is de impact van het gebruik van de tool op beslissingen door rechters onbekend, ook omdat de rechter en de reclasseringsambtenaar altijd zelf een inschatting maken aan de hand van tevens andere informatie. Daarmee is het tevens onduidelijk of sprake is van een eventuele ongelijke behandeling en is er geen informatie over de vraag of het instrument beter presteert dan menselijke beslissers. Vanuit beleidsmatig en praktisch oogpunt rijst op basis hiervan wel de vraag of het, gezien genoemde onzekerheden en risico's, beter is om het instrument voorlopig niet te gebruiken. •

**40.** Derek W. Braverman e.a., 'OxRec model for assessing risk of recidivism: ethics', *The Lancet Psychiatry* 2016/9, p. 808-809.